How many friends can you make in a week?: evolving social relationships in MOOCs over time

Yiqiao Xu
Department of Computer Science
NCSU, Raleigh, NC, USA
yxu35@ncsu.edu

Collin F. Lynch
Department of Computer Science
NCSU, Raleigh, NC, USA
cflynch@ncsu.edu

Tiffany Barnes
Department of Computer Science
NCSU, Raleigh, NC, USA
tmbarnes@ncsu.edu

ABSTRACT

Massive Open Online Courses (MOOCs) are designed on the assumption that good students will help poor students thus offloading the individual support tasks from the instructor to the class. However prior research has shown that this is not always true. Students in MOOCs tend to form distinct sub-communities and their grades are closely correlated with those of their closest peers. That work, however, was only based on analyzing the final social network in a MOOC. In this paper, we study the evolution of these co-performing clusters over time. We explore a longitudinal approach to detect how students form their social connections on the discussion forum and we show that students form close coequal communities early in the course and maintain them over the duration of the course.

Keywords

MOOC, social network analysis, community detection, forum participation

1. INTRODUCTION

One promise of Massive Open Online Courses (MOOCs) is that we can provide high-quality educational content to students around the world at relatively low cost. The broad goal of MOOCs is to scale instruction by allowing expert instructors to provide guidance to hundreds or even thousands of students at a time. Such large-scale education has the potential to be revolutionary both for individual students and for educational systems. The current generation of MOOCs are designed to achieve this scaling by outsourcing much of the individual support tasks to students. That is, rather than capping enrollment to ensure that the instructor and TAs can support every students' needs, MOOCs provide online forums that encourage students to share common questions and to provide collaborative guidance or to benefit from each others' interactions with the limited support staff. Thus it is tacitly assumed that students will have common issues and that good students will help poor students with course content, assignments, logistics, and other issues. The role of instructors and TAs is then often to *curate* help rather than *authoring* it.

In a prior study Brown et al. examined the formation of communities in a large scale MOOC on Big Data in Education [3]. They extracted social networks from the online course forum and analyzed the connections between students. Contrary to the implicit assumption described above, they found that the social connections were not evenly distributed. Nor did they find that the lower-performing students made persistent connections with their higher-performing peers. Instead they found that the students formed distinct sub-communities and that their performance in the course was strongly correlated with that of their closest neighbors. In followup work, Brown et al. also found that these communities were not aligned with students' shared backgrounds nor were they apparently driven by shared course goals [2]. They further found that these results were stable even after the instructional staff and other highly-connected or *hub* students were factored out. Thus the authors concluded that the pattern of students' social relationships can be used to predict their performance and that interventions which target those social relationships may help students to improve either by selecting good peers or by flagging isolated and poorly-performing groups for individual attention.

That work, however, was limited by the fact that it only used the *final* social network from the course. Thus when evaluating students' performance the authors included all posts and social interactions that had developed over the duration of the course. In order to provide useful guidance during the course and to provide reliable information to instructors, we must show that it is possible to detect these relationships based upon partially-formed networks. In general most students' patterns of help-seeking change over the duration of the course. Students often drop out of courses, particularly MOOCs, or taper off their involvement as they lose interest. Students also face difficulties in courses that may make them scale up their communication as the course becomes more challenging. It may be the case that the network structure will change radically over the course of the class and that any early detection model or instructor dashboard will be erratic, invalid, or simply out of date.

In this work, we expand upon the prior work of Brown et al. by examining the growth of the students' social relationships over time, in the same MOOC. To that end we segmented

the forum data by time and performed a sequential analysis of the evolving social network. Our goal in this work will be to address the following questions: First, are students' social groups stable over time? And if so, how early in the course do these observed grade relationships hold? Second, can we use partial social networks to help inform instructors and students in MOOCs? If the answer to these questions is true then it may be possible to develop effective social intervention systems that could use students' posting behaviors to flag students that need attention, or to generate strategic advice on where or how often to post questions. Section 2 provides some background on social network analysis in education. Section 3 describes the dataset we use in our work. In Sections 4 and 5 we present our analysis and results. And finally in section 6 we present our conclusions and discuss our future work.

2. BACKGROUND

2.1 MOOCs, Forums, Students Performance

According to Seaton et al. most of the time students spend on MOOCs is spent viewing the lecture videos, completing mastery assignments, and reading the discussion forums [21]. Very little time is spent on external or 'off-platform' activities. Thus, the discussion forums provide a rich and useful window into the students' primary course activities. Stahl et al. [24] illustrated how students collaborate to create knowledge through this interaction. They argued that students' forum activities are not only beneficial for the individual discussants but also serve to structure the class as a whole. Each student's activity level varies as does their impact on the course. Huang et al. for example, specifically investigated the behavior of high-volume posters in 44 MOOC-related forums. These 'super-posters' tend to enroll in more courses and generally perform better on average [12]. Moreover, by actively engaging in many conversations, they add to the overall volume of the course discussion and they tend to leave fewer questions unanswered in the forums. They also found that, despite their high output, these super-posters did not act to suppress the activity of other less-active users. Rienties et al. [19] examined the way in which students structure their social interactions online. They found that allowing students to self-select collaborators in a MOOC is more conducive to learning than random assignment of partners. In another study, Van Dijk et al[25] found that simple peer instruction is significantly less effective in the absence of a group discussion step, thus reinforcing the importance of a shared class forum.

Prior researchers have also examined the general dynamics of the student forums. Boroujeni et al. examined the relationship between students' temporal patterns, discussion content and social structures emerging from the forums [23]. They found that for MOOCs lasting eight weeks, the pace of students' posts remained high during the first 3 weeks and then tapered down gradually until the class ended. They also found that this pattern was affected by the assignment dates and other deadlines as well as the overall volume of the posts in each thread. Furthermore, they tracked the network attributes over time by using one-week network slices based upon a sliding window. The slice for each day of the course (d>6) was built from forum activities during the preceeding 7 days ([d-6, d]). For each network slice, the attributes included node counts, edge counts, average degree, density, etc. They found that, with the exception of density, the attributes decreased over time. Density, ratio of the number of edges in the graph and the number of edges possible, by contrast, increased sharply at the end of course. Zhu et al. explored a longitudinal approach to combine student engagement, performance, and social connections by applying exponential random graph models [29]. They analyzed the relationship between the social networks on a week-by-week basis and they found that students' individual assignment scores were all positively related to being more active in the social network.

Rosé et al.[20] examined students' evolving social interactions in MOOCs using a Mixed-Membership Stochastic Block model which seeks to detect partially overlapping communities. Their specific focus in the analysis was on identifying the students who were most likely to drop out. They found that it was possible to predict whether or not a student would drop out based upon their membership in a community. Students who actively participated in the forums early on in the course were less likely to drop out later on. Moreover, they found that one specific sub-community was much more prone to dropout than the remainder of the class. This suggests that the forum communities do align by stability and thus that social relationships can reflect the students' relative level of motivation as well as their overall experience in the course. This is akin to the 'emotional contagion' model used in the Facebook mood manipulation study by Kramer, Guillroy, and Hancock [16].

Dawson et al. [6] elaborated the use of social networks to provide guidance. They provided feedback to students and instructors based upon the students' *ego-social* network (i.e. their neighborhood). They explored differences in the network composition for low- and high-performing students to identify patterns of behaviours which may influence the students' learning. They found that the ego-social networks of low- and high-performing students had significant differences, and it was possible to identify different types of students based upon their ego-network. They also found that the instructors were equally likely to show up in high-performing students' local networks as in those of the low-performing students. Their results indicated that instructors could adjust their teaching methods based upon this network structure.

2.2 Communities

There has also been prior research specifically on how students connect within sub-communities and with the instructor. Insa et al. showed that in a traditional course (containing both face-to-face lectures and lab sessions), the student's seating position can affect their final grade [13]. They suggested that physical proximity to the instructor increased performance. According to Golder et al., an analysis of students' Facebook messages showed that the students will message one another more often during weekday afternoons than over the weekend [9]. This produced a distinct temporal pattern in their communication and community structure.

The motivation for any student to join a MOOC can vary widely. This can in turn create several distinct classes of participants with their own unique behaviors. Anderson et al., for example, argued that MOOC participants can be

partitioned into 5 distinct categories based on the number of lectures that they watched and on the assignments that they submitted: viewers, solvers, all-rounders, collectors, and bystanders [1]. They also found that the more assignments a student completed and the more lectures that they viewed, the higher their final grade would be. Interestingly, while students who received a 'B' grade showed a small decrease in their homework submissions relative to 'A' students, the amount of time that those students spent watching lectures was substantially lower. In related work by Liu et al. however, the authors found that some of these behavioral differences were consistent with the students' cultural background which may affect not just their motivation but their expectations and habits [18].

Other authors have examined the relationship between students' academic performance and their social network relationships. Eckles et al. used network analysis on survey data to identify at-risk students who were more likely to drop out [7]. Kovanovic et al. analyzed how a student's relative centrality in their social network will affect their academic performance [15]. They found that more central students were typically higher performers than their less-connected peers. Finally, Zhang et al. constructed student social networks based upon the comments and replies that had been posted to the forum [28]. By analyzing the relative in- and out-degree of the vertices, they were able to identify a small amount of users who answered a large proportion of the questions. This allowed them to find key students in the course.

2.3 Student Behaviours

In their analysis of student behaviors, Anderson et al. found that the number of students who watched lecture videos and finished assignments decreased over the duration of the course, suggesting that some students changed their minds about the class or simply changed their habits during it [1]. Ye et al. performed a similar study, in which they examined a 10-week computer science MOOC [27]. At the end of week 4, 60% of the students who had only watched lectures but had not participated in other ways had dropped out of the course, while only 20% of the students who had submitted assignments and completed quizzes along with viewing had done so.

Given that a large number of MOOC registrants in a given course drop [1, 27], studying the causes of this dropout and preventing it is an important issue. Kloft et al. sought to predict dropout behaviors in a 12-week course based upon the students' click-stream data using a Support Vector Machine [14]. They identified two peak dropout points, one during the first two weeks of the course, and the second at the end of weeks 11 and 12. Students were unlikely to drop in the middle of the course and thus if they made it through the early stages and the final crunch then they would likely complete. Halawa et al. used a specialized definition of drop out as a student being absent from the course for more than 1 month or if they viewed less than half of the lecture videos [10]. With this definition they found that the percentage of students absent from the course sharply decreased from 36.4% to 13.8% after week 3. Hoskins , by contrast, focused exclusively on quizzes as performance-based indicators. They provided a web dashboard for students to self-assess their performance. By comparing students' self-assessments with their grades they found that low performing students tended to drop out more than their higher-performing peers [11].

Unlike the prior studies of students' performance on MOOCs we constructed a temporal social network structure to examine how and when MOOC students established their social connections with differently-performing peers, how their social connections changed over time, and the correlation between these community connections and their intermediate and final performance. We found MOOC students formed their social structures early in the course and that these relationships are stable over time.

3. DATA SET

In this study we used data from a 2013 course on "Big Data in Education" that was offered by the Teachers College at Columbia University and hosted on the Coursera platform. This was an 8-week course that was designed to cover all of the requisite material for a single-semester graduate-level course on Educational Data Mining (EDM) and Big Data analysis in education. This included studying core methods such as student modeling and introducing students to basic data collection and data analysis techniques such as logging and visualization. This iteration of the MOOC ran from October 24, 2013 to December 26, 2013. The course itself was structured around weekly lecture videos and individual assignments or quizzes which contributed to the students' final grade. The weekly assignments were structured around data analysis tasks with students being tasked with conducting some analysis discussed in class and then answering numeric or multiple-choice questions about it. The students were required to complete each assignment within two weeks of its being given out. They were also given up to three attempts per assignment.

The course had a total enrollment of over 48,000 students, but a much smaller number of active participants. 13,314 students watched at least one video while 1,242 watched all of them. A total of 1,380 students completed at least one assignment, and 778 made at least one post or comment in the forum. Of those students who made posts, 426 completed at least one class assignment A total of 638 students completed the online course and received a certificate (meaning that some were able to earn a certificate without participating in forums at all). In order to receive a certificate students were required to earn an overall grade average of 70% or above on the assignments [26].

4. METHODS

We began our analysis by clustering the count of students' submissions for each assignment by date in order to understand when students completed their assignments and how the submission patterns might indicate their working habits. Unsurprisingly the assignment submissions peaked right before each due date with few if any late submissions. To make our analysis consistent we broke the 8-week course into 2-week chunks and we split our analysis at weeks 2 (start), 4 (midterm), 6 (third quarter), and 8 (final). This decision was based upon the fact that students worked across weeks, and on prior literature that pegged the two- and four-week boundaries as crucial times for dropout (e.g. [14, 27]).

This partitioning yielded four distinct datasets representing the cumulative forum discussion up to that point in the class. We extracted a social network from each of these datasets using the same approach applied by Brown et al. [3, 2]. In this approach we generated a raw social network for the course where each node represents a single participant (student, TA, or instructor). We then labeled the student nodes with their cumulative performance up to the specified time step. Thus, the week 2 dataset was labelled using their cumulative performance up to the end of week 2. The Coursera forums operate as standard threaded forums. Users have the ability to start new threads by making an initial post. They can also add posts to the end of an existing thread or add a specific reply below a given post.

In order to build social network from the discussion forum, we treated participants as nodes and their communications as edges. More specifically, for each comment in a thread, we added a directed arc from the author's node to nodes representing the author of each comment that precedes it in the thread, with the exception of self-loops. So all of the contributors to a thread, including the originator, will be connected to one another. This approach is based upon the assumption that students read the thread *before* contributing to it and that a post represents a contribution to the whole conversation. The average length of each thread in our dataset was seven posts. Thus we treat each reply as evidence of an *implicit social connection* between the individual author and their conversational peers. Such implicit social relationships have been explored in the context of recommender systems to detect strong communities of researchers [4]. The resulting networks form a multigraph with each edge representing a single communicative act. As our goal is to focus on social relationships we then modified this graph by eliminating all isolated nodes, and by collapsing the parallel edges to produce a weighted undirected simple graph representing connections between students.

In addition to analyzing the connections between students, we also sought to analyze the impact of the instructional staff and the active hub students on their social structure. We therefore generated three different graphs for each of the datasets: *ALL* which is the complete graph with all non-isolated nodes; *Student*, which eliminates the instructional staff; and *No Hub*, which removes both the instructional staff and the highly active 'hub' students. Since MOOCs are an at-will course students often drop out and we cannot always distinguish intentional dropouts from unintentional failure. In one typical dataset, for example, more than 80% of the students received a grade of 0 [1]. Therefore we also constructed graphs for students with and without students who received a grade of 0. While it is true that the final grade is only accessible at the end of the course we do not believe that this limits the generality of our results. By identifying features that are consistent with 0 performance we can develop predictive models that will work in real-time.

4.1 Best-Friend Regression

Fire et al. modeled students' social interactions for grade prediction in a traditional classroom [8]. They found that in traditional classes the students' grades are closely correlated with those of their closest neighbor or "best friend". That research was based upon self-reported relationship data, but Brown et al. were able to show that it also applied in an online context [2]. In that analysis they used the weighted network to identify each students' "Best Friend" (BF) or closest peer by connections. They then showed that the same result held for this network structure as well.

4.2 Community Detection

We applied the Girvan-Newman algorithm to find social clusters within our graph. In order to identify the ideal number of clusters we used the "natural cluster number" approach described in [3]. That approach is based upon the modularity score of candidate clusters. Given a graph that has been clustered into sub-communities, the modularity of the graph is measured by the ratio of intra-cluster to inter-cluster connections, that is, how strongly individual students are associated with their cluster associates relative to the rest of the class. Graphs with high modularity have very strong within-cluster connections and relatively sparse connections across the groups. As the graphs are partitioned into smaller and smaller communities the modularity score will grow rapidly until we reach an inflection point or a point of diminishing returns at which point each additional sub-cluster makes little difference to or even reduces the modularity score. In the natural cluster approach, we iteratively cluster the graph into higher numbers of communities and plot the modularity score over number of clusters. We then examine this curve to find the inflection point and use that value. This is an exploratory approach similar to exploratory Principal Components Analysis.

4.3 MOOCs, Forums, Student Performance

In MOOCs, the class forum is typically the only official way for students to communicate with the instructors and with each other. Thus, their activities on the forum represent a mostly-complete record of their communicative actions and it represents the best record of their questions and interests. So the dynamic of student forum activities represent their real-time learning status. In order to investigate the dynamics of the students' forum activities and their relationship with the students' social networks, we extracted the number of posts and comments, the number of forum users (who wrote posts) and the number of threads added on a biweekly basis. We then analyzed the numbers in each two-week pair to find the scale of the social network in each case. We also explored how the social aspects of the discussion forum changed over time, by calculating density, degree, average path, diameter and other basic metrics. These network attributes represent the evolving network structure. Furthermore, we compared the scale of the dynamic networks and the network structures to determine when the social networks stabilized. Finally, we analyzed the average number of changes in the neighbors for each student to learn how students selected their communities biweekly.

5. RESULTS & DISCUSSION

Table 1 shows the order (number of nodes) and size (number of edges) of the graphs that we obtained at each cutoff point. While the graphs grew monotonically in order and size over the duration of the course, most of the connections between the students were already established by the end of week two. That is, the basic network structure, if not its weight, was set early on.

Table 1: Graph order and size for each cutoff.

	Nodes	Edges	Comments
Week2	645	14,050	2,472
Week4	693	15,142	3,231
Week6	725	16,346	3,833
Week8	754	17,004	4,260

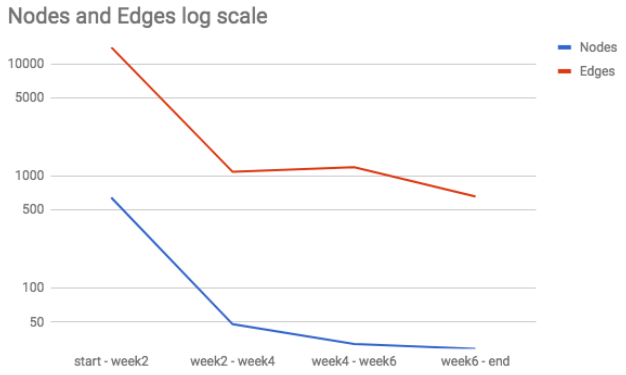

Figure 1: New participating students and connections every two weeks

At the end of the course, there were 55,179 registered users, yet the final course graph contained only 754 participants, 751 of whom were students with 1 instructor and 2 Teaching Assistants. Additionally, 304 of the 751 students obtained a zero grade at the end of the course while 447 received non-zero grades. Some of the forum participants did not complete any assignments but still chose to discuss the course topics with others. By the same token, some of the students who completed work in this course did not participate in the forum at all. There were 1,381 students who received a non-zero final grade; 934 of these did not post in the forum, while 304 zero final grade students did. It is conceivable that when the students met with problems, they chose to ask questions online, but participation in the course forum was not a necessary condition for completion.

Figure 1 shows the number of new participants and new connections added into the social network every two weeks. We applied log scale for the y-axis to make the chart more readable. As these results illustrate almost all students and instructors had established their connections in this course by the end of week two and only a few new connections were made after that time. Additionally, the total number of posts/comments made was 4,260; 2,472 of them (or 58%) had been made at the end of week 2. In our later analysis, we defined a distinct type of 'social connection' post, which includes student-initiated introductions to the class as well as attempts to set up general social connections via Facebook groups, Linkedin links, or other mechanisms. As a results, we collected 182 'social connection' type of student posts. However, even if we discount those 'introduce yourself' comments, it still shows that most of the posting activity happened at the beginning of the course. One potential explanation for this is that the students, particularly those who did not plan to obtain a certificate, did most of their work early and subsequently lost interest. Or, some of the students worked in spurts and did not fit the schedule over time. An ongoing analysis of the forum content has shown that a number of the posts are also about early issues such as course logistics and software, problems which may be less relevant later on. Irrespective of the cause, the social structure is well established early enough that information based upon it can be used to advise students before it is too late.

5.1 Best-Friend Regression

As part of our analysis we also replicated the Best-Friend comparison used by Brown et al. Here we identified each student's closest neighbor in the course, ignoring teaching staff, and we calculated a direct correlation between their grades and those of their best friends. Because the data was non-normal we used Spearman's Rank Correlation Coefficient (ρ), a non-parametric measure of association [22, 5]. Our results are shown in Table 2. Because week 8 is the last week of the course, the intermediate grade is the final grade.

Table 2: Correlation and p-values for Best Friends analysis.

	intermediate grade		final grade	
	ρ	p	ρ	p
Week2	0.25	< 0.001	0.27	< 0.001
Week2_non0	0.086	0.12	0.093	0.08
Week4	0.313	< 0.001	0.339	< 0.001
Week4_non0	0.145	0.005	0.158	0.002
Week6	0.42	< 0.001	0.437	< 0.001
Week6_non0	0.25	< 0.001	0.295	< 0.001
Week8	NA	NA	0.44	< 0.001
Week8_non0	NA	NA	0.29	< 0.001

As shown in Table 2, the students' grade and their best friends' grades, both final and intermediate grades for each bi-week, were strongly correlated, ρ was high, and significant $p < 0.001$. However, the correlation ws affected by the clusters of 0 grade students. After removing these students, the correlations did not hold at a statistically-significant level until the middle of the course. After week four, we found a moderate correlation, $\rho = 0.295$, $\rho = 0.25$, $\rho = 0.29$ and $p < 0.001$. Thus, the relationship between students' grades and those of their best friends were consistent from the traditional face-to-face class to MOOC but not immediately. Our results show that MOOC students, except those who did not submit any assignments, performed similarly to their closest peers.

5.2 Community Detect

Figure 2 provides an example of the modularity curves both with and without zero-score students. We selected natural cluster numbers by finding the inflection points for modularity score. Table 3 shows the selected number of natural clusters based on each week's intermediate grade and table 4 shows the number of clusters based upon the final grade. From table 3 and table4, we found the maximum modularity score for clusters decreases over time. As the modularity score is designed to measure the cleanliness of dividing the network into clusters, these results indicate that the connections between the individual students become more sparse

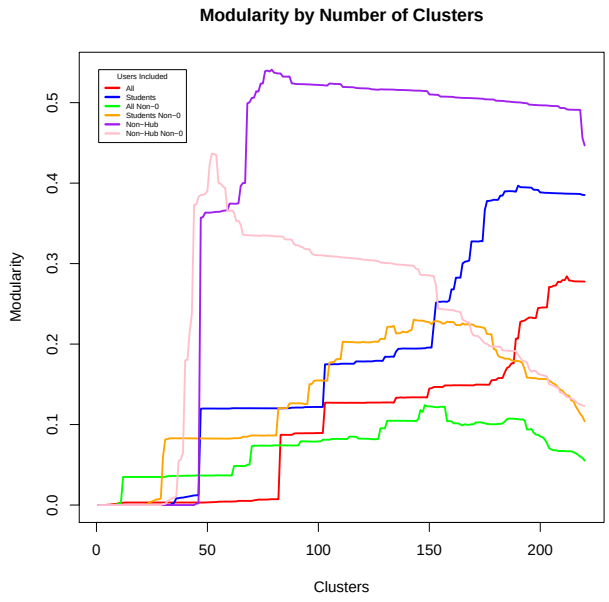

Figure 2: Modularity by Number of Clusters for Week 8

over time while the connections between the clusters of students become more dense as the course progresses.

Interestingly, the curves for the ALL and Hub Student graph are extremely similar, which indicates that hub-students were those who kept a close connection with instructor and TAs. As we anticipated, the non-zero students are the largest group of students. The social network graph shows that many of the zero-score students were only connected with other zero-score students which supports our argument that poor performing students are likely to connect with others at the same performance level.

To assess cluster stability, we also calculated student-centric cluster similarity metrics for the graphs. Tables 5 and 6 show the average number of neighbors that each student loses, gains, or retains in their cluster from week to week. That is, it shows how many former friends are now in a different group, how many new friends are added, and how many stay the same. These figures are shown for weeks 2-4, 4-6, and 6-8 for the all but the no-hub graphs. We excluded the no-hub graphs from our analysis because the models were constructed week by week, the specific hub students did change over time(22% hub group changed from week2 to week8). We also generated the metrics for the social networks based upon the final grades and the weekly cumulative grades. As the tables illustrate, the clusters lost members in each week with the losses being highest in the jump from week 2 to week 4, when the network is still growing quickly. In the later weeks, the losses were smaller, particularly in weeks 4-6. And, for all but the All_NonZero graph, the students gained few new neighbors, with most of the neighbors being retained. As discussed above, the number of clusters increased as the course went on. As these tables indicate

Table 3: Modularity and number of clusters for each graph with intermediate grade

Graph Type	Week2	Week4	Week6
All	112	177	200
Modularity	0.346	0.327	0.276
All_non0	56	100	121
Modularity	0.276	0.195	0.122
Students	119	129	172
Modularity	0.414	0.419	0.393
Students_non0	63	97	125
Modularity	0.436	0.346	0.266
Nonhub	63	67	69
Modularity	0.590	0.590	0.553
Nonhub_non0	43	41	55
Modularity	0.613	0.490	0.396

Table 4: Modularity and number of clusters for each graph with final grade

Graph Type	Week2	Week4	Week6	Week8
All	112	177	200	212
Modularity	0.346	0.327	0.276	0.284
All_non0	56	135	149	173
Modularity	0.257	0.202	0.161	0.103
Students	119	129	172	184
Modularity	0.414	0.419	0.393	0.390
Students_non0	63	109	130	169
Modularity	0.439	0.351	0.304	0.224
Nonhub	63	67	69	79
Modularity	0.590	0.580	0.553	0.541
Nonhub_non0	43	45	49	52
Modularity	0.570	0.478	0.407	0.437

Table 5: Average Dynamic Cluster Changes with final grades

finalgrade	all			all_non0		
week	lost∗	gain∗	overlap∗	lost	gain	overlap
2-4	11.7	1.75	29.6	8.46	2.63	9.94
4-6	1.75	1.75	28	2.53	17.9	9.3
6-8	1.9	3.27	26.74	9.86	36.3	16.9
	students			students_non0		
week	lost	gain	overlap	lost	gain	overlap
2-4	2.05	9.47	30.7	19.3	2.61	11.4
4-6	9.7	1.55	28.77	3.8	2.96	9.42
6-8	1.64	2.72	27.7	2.72	8.94	9.34

lost: average number of lost neighbors
gain: average number of new neighbors
overlap: average number of the same neighbors

the new clusters were generally subsets of the prior clusters and did not present a remix of the prior neighborhoods. The lone exception was the All_NonZero graph which had substantial gains in weeks 4-6 and 6-8. This suggests that the lurkers and other non-certification-seeking students are an important factor in the stability of the social networks; thus, discarding them has a notable effect. However, more analysis is required to understand just how they engender this stability and just how widely distributed they are in the clusters.

Table 6: Average Dynamic Cluster Changes with Intermediate Grades

intergrade	all			all_non0		
week	lost	gain	overlap	lost	gain	overlap
2-4	11.7	1.75	29.6	14.6	20.2	17.1
4-6	1.75	1.75	28	6.87	50	28.8
6-8	1.9	3.27	26.74	41.7	15.3	37.7
	students			students_non0		
week	lost	gain	overlap	lost	gain	overlap
2-4	2.05	9.47	30.7	20.5	3.2	11.2
4-6	9.7	1.55	28.77	3.28	17.5	10.3
6-8	1.64	2.72	27.7	17.3	8.2	10

5.3 Student Performance & Motivation

According to the social network graph, students clustered into different clusters based on their connections and their performance. In order to examine the grade distribution of each cluster, we applied the Kruskal-Walls(KW) test to evaluate the correlation between clusters and performance. The KW test is a non-parametric rank-based similar to the common Analysis of Variance [17]. The result for each graph shown in table 7 - 8 while the 'F' column value is Chi_square. We can see that for nonzero score students, their performance was highly related with their clustered friends, but when all students are included, the relationship becomes weak. This result supports our hypothesis that students will connect with similar performers, instead of helping poor performing students or learning from good ones [26].

Table 7: KW test with intermediate grade

Graph Type	Week2		Week4		Week6	
	F	P	F	P	F	P
all	207	< 0.001	270	< 0.001	315	< 0.001
all_non0	74	0.04	133	0.07	129	0.25
students	218	< 0.001	228	< 0.001	285	< 0.001
students_non0	55	0.69	118	0.06	142	0.12
nonhub	134	< 0.001	171	< 0.001	182	< 0.001
nonhub_non0	53	0.1	47	0.18	90	0.001

Table 8: KW test with final grade

Graph Type	Week2		Week4		Week6	
	F	P	F	P	F	P
all	210	< 0.001	273	< 0.001	319	< 0.001
all_non0	70	0.19	154	0.1	168	0.12
students	223	< 0.001	239	< 0.001	293	< 0.001
students_non0	80	0.06	127	0.1	164	0.01
nonhub	145	< 0.001	179	< 0.001	190	< 0.001
nonhub_non0	44	0.2	58	0.06	67	0.03

Table 9: Forum attributes over time

Attribute	Week2	Week4	Week6	Week8
Posts	2514	3231	3833	4233
Users	659	707	742	770
Threads	345	460	545	597

Table 9 is representative of the evolution of the forum attributes over the 2 week intervals. The overall number of posts, threads, and users increase over time. From the table, we can see that the increase in the number of posts and threads is stable from course start to end. By the end of week 2, 59.4% of the posts had been added to the data and 57.8% of the threads were started in the course forum. However, considering the number of users, 85.6% of the total forum users showed up by week 2. So, by one quarter of the way through the class, most of the users had already showed up in the forum, but fewer than 60% of posts and thread had been initiated. Table 10 shows that the values of the network attributes don't have clear changes which may indicate that the root social network structure doesn't change after week 2. Thus, the dynamics of the forum attributes are consistent with our findings for the best friends and community analysis over time, that the student forum social network structure will develop as soon as week 2 and will then become stable, with the small communities and best friends only getting stronger.

Table 10: Network attributes over time

Attribute	Week2	Week4	Week6	Week8
Degree	21.783	21.850	22.546	22.552
Density	0.034	0.032	0.031	0.030
Avg_path	2.607	2.535	2.492	2.490
Diameter	7	7	7	7
Connected component	82	88	89	98

6. CONCLUSION

Our goal in this paper was to address the potential utility of social network information to guide students and instructors in MOOCs. As prior work has shown, students' final social network structures, particularly their closest neighbors or "best friends" and their sub-clusters, can be analyzed to predict their performance. However, in order to provide meaningful guidance, or to help students and instructors improve their performance before it is too late, it is necessary to show that we can extract useful information from partially-formed social networks. In this paper we have shown that the structure of the students' social networks can be analyzed to predict their performance even by the second week in the course.

Consistent with the prior literature, we found that students are most closely associated with similarly-performing peers and it is possible to predict students' performance based upon their closest neighbors in the graph. Therefore, good students are not necessarily connecting closely with poorer performers, or spreading their help evenly across the class. These results hold even if we remove the instructional staff, hub students, and zero-grade students from the course.

These results suggest that it could be possible to use forum data to identify isolated students or poorly-performing subcommunities that are in need of help. It might also help provide guidance to students who may not be seeking help from the right places. By identifying students who are not isolated, but who are not necessarily getting help from good peers, we may be able to intervene to not only improve their individual standing but also to improve the (social network) structure of the course as a whole. These results also suggest that we should consider mechanisms to encourage more distributed feedback, such as explicit rewards for peer tutoring.

Interestingly, we found that students' social behaviours are consistent because, while students continue to contribute to

the course over time, the social structure of the course is established relatively early. More than half of the forum posts are made in the first two weeks of class. And few students begin to participate on the forum after that point. It is not the case that we have a dynamic graph which can be analyzed differently at each stage. Rather, it appears that the basic structure of the social relationships are fixed early and then only grow stronger over time. While more analysis is required to determine why this occurs, it suggests that students' initial impressions or choices have a strong impact on their performance and that interventions which are designed to change those habits may be beneficial. One avenue of research that we are currently pursuing is to analyze the content of the individual posts. If we can detect a change in the nature or structure of the content or of the topics being considered it might help to explain why the students' progress appears to taper off so dramatically. At the same time we plan to experiment with evaluating metrics of this type for blended courses to see if similar dynamic results hold in blended face-to-face and online contexts.

Furthermore, our results indicate that a social network analysis of the discussion forum data brings an unprecedented opportunity for instructors to visualize students' social structures and to form learning networks which allow them to make changes to their teaching plans over time. For nonzero grade students, the correlation between students' grades and their best friends' grades is not reliable during the first 4 weeks of the course. However network features may be useful for early detection of at-risk students. Real-time egonetworks may also explain how low performance is related to connections to other low performing students. This suggests that it may be useful to incentivize high performing students to make connections with lower performing student threads.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1418269: "Modeling Social Interaction & Performance in STEM Learning" Yoav Bergner, Ryan Baker, Danielle S. McNamara, & Tiffany Barnes Co-PIs.

8. REFERENCES

[1] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec. Engaging with massive online courses. In *Proceedings of the 23rd international conference on World wide web*, pages 687–698. ACM, 2014.

[2] R. Brown, C. Lynch, Y. Wang, M. Eagle, J. Albert, T. Barnes, R. Baker, Y. Bergner, and D. S. McNamara. Communities of performance & communities of preference. In *EDM (Workshops)*, 2015.

[3] R. Brown, C. F. Lynch, M. Eagle, J. Albert, T. Barnes, R. Baker, Y. Bergner, and D. S. McNamara. Good communities and bad communities: Does membership affect performance? In *EDM (Workshops)*, 2015.

[4] E. Choo, T. Yu, M. Chi, and Y. Sun. Revealing and incorporating implicit communities to improve recommender systems. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 489–506. ACM, 2014.

[5] P. Dalgaard. *Introductory Statistics with R*. Springer Verlag New York Inc., 2002.

[6] S. Dawson. âĂŸseeingâĂŹthe learning community: An exploration of the development of a resource for monitoring online student networking. *British Journal of Educational Technology*, 41(5):736–752, 2010.

[7] J. E. Eckles and E. G. Stradley. A social network analysis of student retention using archival data. *Social Psychology of Education*, 15(2):165–180, 2012.

[8] M. Fire, G. Katz, Y. Elovici, B. Shapira, and L. Rokach. Predicting student examâĂŹs scores by analyzing social network dataanderson. In *International Conference on Active Media Technology*, pages 584–595. Springer, 2012.

[9] S. A. Golder, D. M. Wilkinson, and B. A. Huberman. Rhythms of social interaction: Messaging within a massive online network. *Communities and technologies 2007*, pages 41–66, 2007.

[10] S. Halawa, D. Greene, and J. Mitchell. Dropout prediction in moocs using learner activity features. *Experiences and best practices in and around MOOCs*, 7, 2014.

[11] S. L. Hoskins and J. C. Van Hooff. Motivation and ability: which students use online learning and what influence does it have on their achievement? *British journal of educational technology*, 36(2):177–192, 2005.

[12] J. Huang, A. Dasgupta, A. Ghosh, J. Manning, and M. Sanders. Superposter behavior in mooc forums. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 117–126. ACM, 2014.

[13] D. Insa, J. Silva, and S. Tamarit. Where you sit matters how classroom seating might affect marks. In *Proceedings of the 2016 ACM Conference on Innovation and Technology in Computer Science Education*, pages 212–217. ACM, 2016.

[14] M. Kloft, F. Stiehler, Z. Zheng, and N. Pinkwart. Predicting mooc dropout over weeks using machine learning methods. In *Proceedings of the EMNLP 2014 Workshop on Analysis of Large Scale Social Interaction in MOOCs*, pages 60–65, 2014.

[15] V. Kovanovic, S. Joksimovic, D. Gasevic, and M. Hatala. What is the source of social capital? the association between social network position and social presence in communities of inquiry. 2014.

[16] A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

[17] W. H. Kruskal and W. A. Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[18] Z. Liu*, R. Brown*, **C. F. Lynch**, T. Barnes, R. Baker, Y. Bergner, and D. Mcnamara. Mooc learner behaviors by country and culture; an exploratory analysis. In T. Barnes, M. Chi, and M. Feng, editors, *Proceedings of the 2016 Conference on Educational Data Mining*. International Educational Data Mining Society, 2016.

[19] B. Rienties, P. Alcott, and D. Jindal-Snape. To let

students self-select or not: that is the question for teachers of culturally diverse groups. *Journal of Studies in International Education*, 18(1):64–83, 2014.

[20] C. P. Rosé, R. Carlson, D. Yang, M. Wen, L. Resnick, P. Goldman, and J. Sherer. Social factors that contribute to attrition in moocs. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 197–198. ACM, 2014.

[21] D. T. Seaton, Y. Bergner, I. Chuang, P. Mitros, and D. E. Pritchard. Who does what in a massive open online course? *Commun. ACM*, 57(4):58–65, Apr. 2014.

[22] P. Sedgwick. Spearman's rank correlation coefficient. *BMJ: British Medical Journal (Online)*, 349, 2014.

[23] M. Shirvani Boroujeni, T. Hecking, H. U. Hoppe, and P. Dillenbourg. Dynamics of mooc discussion forums. In *7th International Learning Analytics and Knowledge Conference (LAK17)*, number EPFL-CONF-223718, 2017.

[24] G. Stahl, T. Anderson, and D. Suthers. Computersupported collaborative learning: An historical perspective, 2006. *Cambridge handbook of the learning sciences*, pages 409–426, 2006.

[25] L. Van Dijk, G. Van Der Berg, and H. Van Keulen. Interactive lectures in engineering education. *European Journal of Engineering Education*, 26(1):15–28, 2001.

[26] Y. Wang and R. Baker. Content or platform: Why do students complete moocs? *Journal of Online Learning and Teaching*, 11(1):17, 2015.

[27] C. Ye and G. Biswas. Early prediction of student dropout and performance in moocss using higher granularity temporal information. *Journal of Learning Analytics*, 1(3):169–172, 2014.

[28] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on World Wide Web*, pages 221–230. ACM, 2007.

[29] M. Zhu, Y. Bergner, Y. Zhang, R. Baker, Y. Wang, and L. Paquette. Longitudinal engagement, performance, and social connectivity: a mooc case study using exponential random graph models. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, pages 223–230. ACM, 2016.